Name:

AMATH 515 Homework Set 1

Due: Monday Jan 23rd, by midnight.

(1) Let $g : \mathbb{R}^m \to \mathbb{R}$ is a twice differentiable function, $A \in \mathbb{R}^{m\times n}$ any matrix, and h is the composition $g(Ax)$, then we have two simple generalizations of the chain rule that combine linear algebra with calculus:

$$\nabla h(x) = A^T \nabla g(Ax)$$

and

$$\nabla^2 h(x) = A^T \nabla^2 g(Ax) A.$$

(a) Show what happens when you apply the above chain rules to the special case

$$h(x) = g(a^T x)$$

where a is a vector.

Suppose $A = a^T$ where $a \in 1 \times m$, then

$$\begin{aligned}
\nabla h(x) &= \nabla g(a^T x)\\
&= \nabla g(Ax)\\
&= A^T \nabla g(Ax) \qquad (1)\\
&= a \nabla g(Ax)\\
&= a \nabla g(a^T x)
\end{aligned}$$

(b) Compute the gradient and hessian of the regularized logistic regression objective:

$$\left(\sum_{i=1}^{n} \log(1 + \exp(a_i^T x)) - b^T Ax\right) + \lambda \|x\|^2$$

where a_i denote the rows of A.

Let $h(x)$ denote our given function. Suppose $g(x) = \log(1 + \exp(x))$, then

$$\begin{aligned}
g(a_i^T x) &= \log(1 + \exp(a_i^T x))\\
\nabla g(a_i^T x) &= a_i \nabla g(a_i^T x) \qquad (1)\\
&= a_i \frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)}
\end{aligned}$$

And the hessian of $g(a_i^T x)$ is,

(2)
$$\nabla^2 g(a_i^T x) = a_i a_i^T \frac{\exp(a_i^T x)}{(1+\exp(a_i^T x))^2}$$

Taking the second term of the summation,

$$\nabla(b^T Ax) = bA^T$$
$$\nabla^2(b^T Ax) = 0$$

Taking the last term,

$$\begin{aligned}
\nabla\lambda\|x\|^2 &= \nabla\lambda(x_1^2 + x_2^2 + x_3^2 + ... + x_n^2) \\
&= 2\lambda x \\
\nabla^2\lambda\|x\|^2 &= \nabla 2\lambda x \\
&= 2\lambda I
\end{aligned}$$

Finally, putting our terms together

$$\begin{aligned}
\nabla h(x) &= \sum_{i=1}^{n} a_i \frac{\exp(a_i^T x)}{1+\exp(a_i^T x)} - bA^T + 2\lambda x \\
\nabla^2 h(x) &= \sum_{i=1}^{n} a_i a_i^T \frac{\exp(a_i^T x)}{(1+\exp(a_i^T x))^2} + 2\lambda I
\end{aligned}$$

(c) Compute the gradient and hessian of the regularized poisson regression objective:

$$\left(\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax\right) + \lambda\|x\|^2$$

where a_i denote the rows of A.

Let $h(x)$ denote the poisson regression objective. Suppose $g(x) = \exp(x)$. By rule 1,

(1)
$$\begin{aligned}
\nabla g(a_i^T x) &= \nabla \exp(a_i^T x) \\
&= a_i \exp(a_i^T x) \\
\nabla^2 g(a_i^T x) &= \nabla a_i \exp(a_i^T x) \\
&= \nabla a_i a_i^T \exp(a_i^T x)
\end{aligned}$$

Using our results from (b) and putting our terms together,

$$\nabla h(x) = \sum_{i=1}^{n} a_i \exp(a_i^T x) - bA^T + 2\lambda x$$
$$\nabla h^2(x) = \sum_{i=1}^{n} a_i a_i^T \exp(a_i^T x) + 2\lambda I$$

(d) Compute the gradient and hessian of the regularized 'concordant' regression objective

$$\|Ax - b\|_2 + \lambda\|x\|_2.$$

Give conditions on a point x that ensure the gradient and Hessian exist at x.
Let $h(x)$ denote our function. Suppose $g(k) = ||k - b||$, then

$$\nabla g(k) = \frac{k-b}{||k-b||}$$
$$\nabla g(Ax) = A^T \frac{Ax-b}{||Ax-b||}$$
$$\nabla^2 g(Ax) = A^T \left(\frac{1}{\|Ax-b\|} - \frac{(Ax-b)}{\|Ax-b\|^2} \right) A$$

Similarly,

$$\nabla f(x) = \lambda \frac{x}{||x||}$$
$$\nabla^2 f(x) = \frac{1}{||x||} - \frac{x}{||x||^2}$$

Putting the above values together,

$$\nabla h(x) = A^T \frac{Ax-b}{||Ax-b||} + \lambda \frac{x}{||x||}$$
$$\nabla^2 h(x) = A^T \left(\frac{1}{||Ax-b\|} - \frac{(Ax-b)}{\|Ax-b\|^2} \right) A + \frac{1}{||x||} - \frac{x}{||x||^2}$$

Note that the values of $||x||$ and $||Ax - b||$ should be different from 0.

(2) Show that each of the following functions is convex.

(a) Indicator function to a convex set: $\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ \infty & \text{if} \quad x \notin C. \end{cases}$

Suppose $y, z \in C$. Let x be a convex combination of y and z given by $x = \lambda y + (1 - \lambda)z$, where $\lambda \in [0, 1]$. For the indicator function to be convex, we want $\delta_C(x) \le \lambda\delta_C(y) + (1 - \lambda)\delta_C(z)$. Since $\delta_C(y), \delta_C(z) = 0$, our indicator function gives 0 for $x \in C$, we get $0 = 0$. Hence the indicator function is convex.

(b) Support function to any set:

$$\sigma_C(x) = \sup_{c \in C} c^T x.$$

$$\begin{aligned}
\sigma_C(x) &= \sigma_C(\lambda x_1 + (1-\lambda)x_2) \\
&= \sup_{c \in C} c^T(\lambda x_1 + (1-\lambda)x_2) \\
&\leq \lambda \sup_{c \in C} c^T x_1 + (1-\lambda) \sup_{c \in C} c^T x_2 \\
&= \lambda \sigma_C(x_1) + (1-\lambda)\sigma_C(x_2)
\end{aligned}$$

(c) Any norm (see Chapter 1 for the definition of a norm).
For all points x, y, the norm in the vector space holds $||x|| \geq 0$, $||\alpha x|| = \alpha ||x||$ and satisfies the triangle inequality.

(3) Convexity and composition rules. Suppose that f and g are $\mathcal{C}^2$ functions from $\mathbb{R}$ to $\mathbb{R}$, with $h = f \circ g$ their composition, defined by $h(x) = f(g(x))$.

(a) If f and g are convex, show it is possible for h to be nonconvex (give an example). Give additional conditions that ensure the composition is convex.

Proof by counter example: let $f(x) = x^2$ and $g(x) = x^2 - 1$. Then $h(x) = f(g(x)) = (x^2-1)^2 = x^4 - 2x^2 + 1$. Computing $h'(x) = 4x^3 - 4x$ and $h''(x) = 12x^2 - 4$. For $x = 0$, $h''(x) < 0$ hence it is possible for $h(x)$ to be nonconvex. $f(x)$ should be a non-decreasing function for $h(x)$ to be convex.

(b) If f is convex and g is concave, what additional hypothesis that guarantees h is convex?

$f(x)$ should be non-increasing function to guarantee that $h(x)$ is a convex function.

(c) Show that if $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ affine, then h is convex.

Since g is affine, it can be represented by $g(x) = Ax + b$. Let $h(x) = f(g(x))$. For any $x, y \in \mathbb{R}^n$. Computing,

$$\begin{aligned}
h(\lambda x + (1-\lambda)y) &= f(g(\lambda x + (1-\lambda)y)) \\
&= f(\lambda Ax + (1-\lambda)Ay + b) \\
&\leq \lambda f(Ax + b) + (1-\lambda)f(Ay + b) \\
&= \lambda h(x) + (1-\lambda)h(y)
\end{aligned}$$

hence by definition of convex function, h is convex.

(d) Show that the following functions are convex:

(i) Logistic regression objective: $\sum_{i=1}^{n} \log(1+\exp(a_i^T x)) - b^T Ax$

Suppose $f(x) = \log(1+\exp(x))$. $g(x)$ is a convex since $f''(x) \geq 0 \forall x \in \mathbb{R}$. $a_i^T x$ is an affine function and let $g(x) = a_i^T x$. Hence by (c), fog is convex function. Convexity is also preserved in matrix operation. Therefore, logistic regression objective is convex.

(ii) Poisson regression objective: $\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax$.

Let $f(x) = exp(x)$. Then $f''(x) = e^x \geq 0, \forall x \in \mathbb{R}$. Hence $f(x)$ is a convex function. Similar to (i), $a_i^T x$ is an affine representation and let $g(x) = a_i^T x$. fog is a convex function by (c). Again convexity is preserved under linear matrix operations. Hence Poisson regression objective is convex.

(4) A function f is *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y), \quad \lambda \in (0,1).$$

(a) Give an example of a strictly convex function that does not have a minimizer.

$$f(x) = e^x$$

(b) Show that a sum of a strictly convex function and a convex function is strictly convex.
Assume that $f(x)$ is a strictly convex function and $g(x)$ is a convex function, then computing their sum

$$\begin{aligned}
f(x) + g(x) &= f(\lambda x_1 + (1-\lambda)x_2) + g(\lambda x_1 + (1-\lambda)x_2) \\
&< \lambda f(x_1) + (1-\lambda)f(x_2) + g(\lambda(x_1) + (1-\lambda)(x_2)) \\
&\leq \lambda f(x_1) + (1-\lambda)f(x_2) + \lambda g(x_1) + (1-\lambda)g(x_2) \\
&= \lambda h(x_1) + (1-\lambda)h(x_2)
\end{aligned}$$

hence their sum is a strictly convex function.

(c) Characterize all solutions to the problem

$$\min_x \frac{1}{2}\|Ax - b\|^2$$

Setting $\nabla f(x) = 0$, we have $A^T(Ax - b) = 0$. Hence any x that satisfies $A^T Ax = A^T b$ will be a solution to the problem.

(5) A function f is β-smooth when its gradient is β-Lipschitz continuous.

(a) Find a global bound for β of the least-squares objective $\frac{1}{2}\|Ax - b\|^2$.

$$\beta = \lambda_{max}(A^T A)$$

(b) Find a global bound for β of the regularized logistic objective

$$\sum_{i=1}^{n} \log(1+\exp(\langle a_i, x\rangle)) + \frac{\lambda}{2}\|x\|^2.$$

$$\beta \leq \frac{1}{4}\lambda_{max}(A^T A)$$

(c) Do the gradients for Poisson regression admit a global Lipschitz constant?

No, since Poisson regression does not have a single β.

(6) Please complete the coding homework (starting with the notebook uploaded to Canvas).